

# CAPACITY AND DELAY TRADEOFF OF SECONDARY CELLULAR NETWORKS WITH SPECTRUM AGGREGATION

Lingyu Chen<sup>1</sup>, Chen Liu<sup>1</sup>, Xuemin Hong<sup>1</sup>, Cheng-Xiang Wang<sup>2</sup> *Fellow, IEEE*, John Thompson<sup>3</sup>, *Fellow, IEEE*, and Jianghong Shi<sup>1</sup>

<sup>1</sup>Department of Communications Engineering,  
School of Information Science and Technology,  
Xiamen University, Xiamen 361005, Fujian, P.R.China.  
Email: {chenly, chenl, xuemin.hong, shijh}@xmu.edu.cn

<sup>2</sup>Institute of Sensors, Signals and Systems,  
School of Engineering and Physical Sciences,  
Heriot-Watt University, Edinburgh, EH14 4AS, UK.  
Email: cheng-xiang.wang@hw.ac.uk

<sup>3</sup>Institute for Digital Communications,  
School of Engineering,  
University of Edinburgh, Edinburgh, EH9 3JL, UK.  
Email: john.thompson@ed.ac.uk

arXiv:1612.08778v1 [cs.NI] 28 Dec 2016

## Abstract

Cellular communication networks are plagued with redundant capacity, which results in low utilization and cost-effectiveness of network capital investments. The redundant capacity can be exploited to deliver secondary traffic that is ultra-elastic and delay-tolerant. In this paper, we propose an analytical framework to study the capacity-delay tradeoff of elastic/secondary traffic in large scale cellular networks with spectrum aggregation. Our framework integrates stochastic geometry and queueing theory models and gives analytical insights into the capacity-delay performance in the interference limited regime. Closed-form results are obtained to characterize the mean delay and delay distribution as functions of per user throughput capacity. The impacts of spectrum aggregation, user and base station (BS) densities, traffic session payload, and primary traffic dynamics on the capacity-delay tradeoff relationship are investigated. The fundamental capacity limit is derived and its scaling behavior is revealed. Our analysis shows the feasibility of providing secondary communication services over cellular networks and highlights some critical design issues.

## I. INTRODUCTION

The capacity of a cellular radio access network (RAN) is fundamentally limited by the density of base stations (BSs), system bandwidth, and spectrum efficiency. Once a particular network is rolled out, its maximum capacity is relatively stable. The traffic load, on the other hand, changes dynamically across space and time. Because the capacity of a cellular network is planned to accommodate peak traffic demand, redundant capacity is unavoidable due to traffic fluctuations. Measurements campaigns (e.g., [1]) have shown that redundant capacity is a pervasive problem, which results in low utilization and cost-effectiveness of network capital investments.

Measurement also revealed that the major cause of mobile traffic is multi-media consumption [2], which includes different types of communication services. The first type is streaming services that are delay-sensitive but loss-tolerant. Typical applications include voice over IP and video conferencing. The second type is elastic traffic services that are delay-tolerant but loss-sensitive. Typical applications include web browsing and file transfer. In practice, there is no crucial difference in the delay constraints of streaming and elastic traffic. However, the emergence of new applications such as proactive caching [3]–[5] brings a third type of traffic that has crucial difference from the first two types. Proactive caching systems are able to push content and cache them closer to end users, taking advantage of the fact that content demand is predictable and large cache space is becoming affordable. The traffic generated by proactive caching is called ultra-elastic traffic as the delay constraint is very relaxed (because user request happens much later) and the traffic demand is very flexible (because caching is transparent and opportunistic). It has been envisioned that the redundant capacity in cellular networks can be exploited to deliver ultra-elastic traffic as secondary traffic, which coexists with other higher priority traffic in the same cellular network [6]. Such a background motivates our study in this paper to investigate the fundamental capacity-delay tradeoff of secondary traffic in cellular networks.

Capacity is a key performance metric for cellular networks. It is widely believed that spectrum aggregation and cell densification will become prominent features of future cellular networks [7], [8], leading to a heterogeneous cellular network (HCN) across multiple separated spectrum bands [9]. The capacity and coverage performance of HCN has been studied extensively in the literature [9]–[21] using stochastic geometry. However, most of these studies focus on the physical layer capacity (i.e., spectrum efficiency), which differs from the throughput capacity considered in this paper. The throughput capacity of a user is evaluated with respect to the traffic queueing process of the user, and is defined as the average traffic arrival density (bits/s) that can be accommodated with finite delay. Throughput capacity addresses the higher layer aspects of traffic queueing and is better suited for cross-layer and delay related studies.

Delay is another key performance metric for analyzing cellular network traffic. Traffic behaviors can be modeled at the packet level or the session/flow level. Packet level dynamics are notoriously complicated as the temporal statistics exhibit self-similarity and multi-fractal behavior [22]. On the contrary, session level behaviors justify the convenient assumption of Poisson arrival and can better reflect how the traffic performance is perceived by end users [23]. Multiple modeling frameworks have been used for delay

analysis in cellular systems with spectrum aggregation. The classic framework is queuing theory, which is theoretically mature but inflexible [24], [25]. Discrete/continuous time Markov chain models [26]–[32] are more flexible for describing some practical aspects, but fall short in providing closed-form analytical insights. The concept of local delay was proposed in [33], [34], but it focuses on the average delay and reveals no information about the delay distribution. In this paper, we adapt queueing theory as the framework to study the session level behavior of secondary traffic.

It is well-known that there is a fundamental tradeoff between capacity and delay. For mobile ad hoc networks, the capacity-delay tradeoff has been studied extensively using the framework of scaling law analysis [35]–[39]. The methodology and results therein, however, is not applicable to cellular networks. The capacity-delay tradeoff in cellular networks is still under-investigated due to the lack of a well-established analytical framework. Some early attempts [40]–[42] use interference approximation techniques to bound the session level performance of multi-cell networks. However, these works are not compatible with popular stochastic geometry models and only provide loose bounds for the estimation of mean delays. A framework of timely throughput was proposed in [43] and recently adopted for the analysis of HCN in [44], [45]. It assumes that a queuing packet will be dropped if the packet passes a critical delay. This is a useful framework that offers analytical tractability and flexibility in taking into account aspects such as mobility, user association bias, etc. However, this framework is better suited for the study of streaming traffic instead of elastic traffic, and does not provide a detailed characterization of delay distribution. Recent attempts to integrate stochastic geometry and queueing models was reported in [46]–[48]. In [46], the spatial-temporal dependence of a cellular system is captured by some cell-load equations and eventually resolved via static simulations. Although this framework is mathematically rigorous, it lacks the analytical tractability to reveal closed-form insights. In our previous work [47], [48], stochastic geometric and queueing models are combined to study the uplink capacity of hybrid ad-hoc networks with user collaboration. However, these work focused on a different type of network and did not fully address the issue of multi-user access, which is a critical feature of cellular systems. Moreover, [47], [48] did not explain how spatial-temporal dependencies in cellular networks can be decoupled to offer analytical insights. To our best knowledge, full integration of stochastic geometry and queueing theory in cellular networks is an unsolved and challenging problem due to complex coupling of network behavior in spatial and temporal domains [46].

This paper proposes a new framework for the study of capacity-delay tradeoff of secondary/elastic traffic in cellular networks. Our framework integrates stochastic geometry and priority queueing models to offer analytical tractability and the ability to pinpoint delay distributions. The analytical tractability of our framework comes from two aspects. First, instead of trying to work out the exact mapping between spatial and temporal domains, our methodology is to identify a set of critical parameters in both domains and establish the relationships among their first-moment measures (mean values). Second, by focusing on the secondary traffic, we are able to justify certain assumptions and approximations. Our framework is shown to be useful in revealing some analytical insights. Specifically, this paper makes the following contributions.

- Analytical results are derived to characterize the mean delay and delay distribution as functions of per user throughput capacity.
- Analytical results are derived to characterize the fundamental capacity limit in some special cases.
- A concise analytical approximation is obtained to describe how the per user capacity scales with user-BS density ratio.

The remainder of this paper is organized as follows. Section II describes the system model. The overall methodology and some useful approximations are introduced in Section III. The capacity-delay tradeoff and fundamental capacity limit are studied in Sections IV and V, respectively. Section VI provides numerical results and discussions. Finally, conclusions are drawn in Section VII.

## II. SYSTEM MODEL

### A. Secondary access protocol

We consider the downlink of a large scale cellular network that aggregates  $N$  independent frequency bands. BSs operating in the same band are assumed to have homogeneous bandwidth and transmit power denoted by  $W_n$  and  $P_n$ , respectively, where  $n$  ( $1 \leq n \leq N$ ) is the index of bands. A user can operate in one band at a time, but can handover between different bands. Secondary users are assumed to comply with the following access protocol.

- Step 1: Periodically check the buffer of secondary traffic. If the buffer is empty, remain in idle mode. Otherwise turn into active mode and proceed to Step 2.
- Step 2: Randomly select one band and associate with the nearest BS operating in the chosen band. This implies a Poisson-Voronoi cell model, which is a widely used cellular network model.
- Step 3: Evaluate whether the associated BS is vacant (i.e., not occupied by primary traffic) and available for secondary services. If yes, proceed to Step 4, otherwise return to Step 2. The probability that a typical BS in the  $n$ th band is vacant is called “vacant probability” and is denoted by  $\Omega_n$ . This parameter indicates the average load of primary traffic.
- Step 4: Evaluate the link quality with respect to the associated BS. If the signal-to-noise-and-interference ratio (SINR) is large enough to support a transmission rate of  $R$  bits/s, proceed to Step 5, otherwise return to Step 2. Here,  $R$  is the minimum rate requirement of secondary transmission. Such a requirement is imposed because it is desirable to restrict secondary services only to users with high quality links, otherwise secondary services may become inefficient and will result in excessive energy consumption and interference. The probability that a typical user in the  $n$ th band has good link quality is called “coverage probability” and denoted by  $p_{n,v}$ .
- Step 5: Compete with other in-coverage users for multiple access to the same BS. We assume a time-division multiple access (TDMA) scheme for multi-user access, where a band is fully allocated to one user at a time and multiple contending users have equal opportunities to access the band through time sharing. If contention is successful, proceed to Step 6, otherwise return to Step 2. The probability that a in-coverage secondary user in the  $n$ th tier is granted access is called “access probability” and denoted by  $p_{n,a}$ .
- Step 6: Transmit secondary traffic with a fixed rate  $R$  until the buffer is empty. If the buffer is empty, proceed to Step 1. Otherwise if an outage (caused by primary traffic interruption or coverage outage) occurs during transmission, return to Step 2. We note that the transmission rate of secondary service is fixed to  $R$  for simplicity.

For a user to receive secondary service in the  $n$ th band, he should firstly be associated with a vacant BS, secondly have a good coverage, and finally be granted access after multi-user contention. It follows that the service probability  $\varepsilon_n$  is the product of vacant probability  $\Omega_n$ , coverage probability  $p_{n,v}$ , and access probability  $p_{n,a}$ , i.e.,

$$\varepsilon_n = \Omega_n \cdot p_{n,v} \cdot p_{n,a}. \quad (1)$$

### B. Spatial interference model

The spatial layout of BSs operating in the  $n$ th band is modeled by a stationary Poisson Point Process (PPP) in  $\mathbf{R}^2$  with intensity  $\lambda_{b,n}$ , which is a commonly used model in the literature. For analytical tractability, we ignore the case of co-located BSs and assume that the spatial layout of BSs in different bands are independent. The spatial distribution of secondary users are also assumed to follow a stationary PPP in  $\mathbf{R}^2$  with intensity  $\lambda_u$ . Let us consider a typical user in the  $n$ th band, the downlink SINR is a random variable, whose cumulative density function (CDF) has been derived for different types of fading channels [50]. For purposes of clarity and tractability, we consider a representative case in which the path

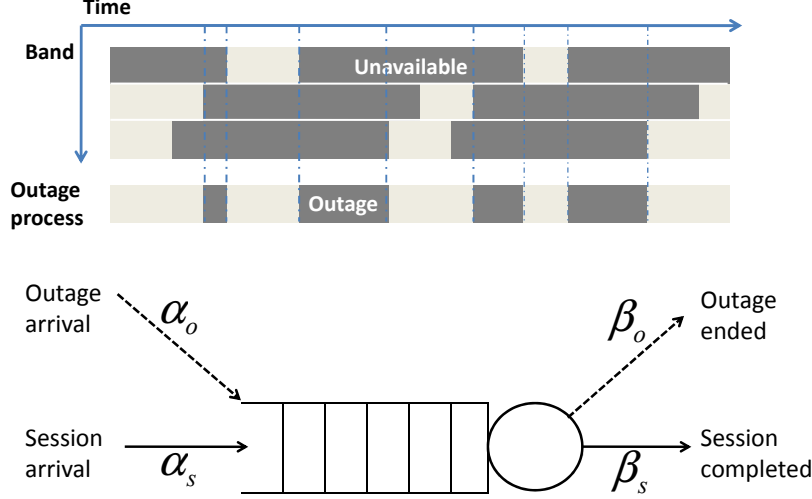


Fig. 1. Priority queuing model of a typical user with secondary traffic and random outage.

loss exponent is 4. The complementary CDF of the user SINR is then given by [50]

$$F_{\gamma,n}(x) = \frac{\pi^{\frac{3}{2}} \lambda_{b,n}}{\sqrt{x/P_n}} e^{\frac{a^2}{\sqrt{2b}}} Q\left(\frac{a}{\sqrt{2x/P_n}}\right) \quad (2)$$

where  $Q(\cdot)$  denotes the  $Q$ -function and

$$a = \lambda_{b,n} \pi [1 + \sqrt{x} \arctan(\sqrt{x})]. \quad (3)$$

If the system is interference limited, which implies that  $P_n$  is sufficiently large and the noise is negligible, (2) can be further simplified to [50]

$$F_{\gamma}^{\text{lim}}(x) = \frac{1}{1 + \sqrt{x} \arctan(\sqrt{x})}. \quad (4)$$

According to the secondary access protocol, a user is in coverage of secondary services if  $W_n \log_2(1 + \gamma_n) \geq R$ , where  $\gamma_n$  denotes the SINR perceived by the user. The coverage probability in the  $n$ th band is therefore given by

$$p_{n,v} = F_{\gamma,n}(2^{R/W_n} - 1). \quad (5)$$

### C. Temporal queuing model

As illustrated in Fig. 1, we model the secondary traffic dynamic as a preemptive priority queue, where the transmission of secondary traffic may be preempted (i.e., immediately interrupted) by outages. An outage can be caused by multiple factors such as primary traffic interruption, bad coverage, and failure in multi-user contention. We assume users can handover between bands with negligible time, hence an outage only occurs when no band is available for secondary services. We propose to model the composite outage effect as a stream of higher priority traffic in the priority queue. The arrival of outage events follows a Poisson process with mean interval  $\bar{\alpha}_o$ . Each outage event contributes to an additive random outage duration denoted by  $\beta_o$ , the mean of which is  $\bar{\beta}_o$ . Let us define

$$\rho_o = \bar{\beta}_o / \bar{\alpha}_o. \quad (6)$$

This parameter represents the fraction of time that a user is in outage and cannot be served by a BS in all bands. It is worth noting that we do not make any particular assumption on the distribution of  $\beta_o$ , i.e., it

can follow an arbitrary form of continuous distribution. This gives our model the flexibility to represent a wide range of outage phenomena.

We consider the secondary traffic behavior at the session level. Users are assumed to have homogeneous incoming traffic of sessions that follow i.i.d. Poisson arrival process with mean interval  $\bar{\alpha}_s$ . Each session carries a file of random size  $L$  to be delivered from the BS to the user. The file size  $L$  follows a general distribution with mean  $\bar{L}$ . The mean throughput capacity of a user is given by

$$C = \bar{L}/\bar{\alpha}_s. \quad (7)$$

Under the assumption of constant transmission rate  $R$ , the transmission time of a session is a random variable  $\beta_s = L/R$ . Let us define

$$\rho_s = \bar{\beta}_s/\bar{\alpha}_s = \bar{L}/(R\bar{\alpha}_s) = C/R. \quad (8)$$

This parameter represents the fraction of time that a user receives transmission from a BS. The file size  $L$  is assumed to follow a general distribution.

The transmission of a secondary session is forced to stop immediately once an outage occurs. Once the secondary service is available again, a session may adopt a ‘resume’ policy to transmit from where it stopped, or adopt a ‘repeat’ policy to retransmit from the beginning. Our paper is restricted to the resume policy, noting that an extension to the repeat policy is straightforward. Based on the above modeling assumptions, the queuing process at a typical secondary user can be readily captured by a M/G/1 two-level priority queuing model with a preemptive resume policy [54]. Such a classic queuing model is fully characterized by the four random variables shown in Fig. 1.

### III. METHODOLOGY AND APPROXIMATIONS

Our system model describes a large scale, dynamic system in the spatial and temporal domains. These two domains are inherently coupled and correlated. Analysis encompassing both domains requires integration of stochastic geometry and queueing models, which is an extremely challenging task. Existing work resorted to static simulation to yield results without revealing much theoretical insight [46]. In this paper, instead of trying to capture the detailed relationships between the spatial and temporal domains, we propose a methodology that connects these two domains by establishing analytical relationships among the first-order statistic measure (i.e., mean values) of some critical parameters. Higher order statistics are not our primary concern, therefore we use the flexible model of M/G/1 queue (with general distributions) to offer sufficient flexibility to represent a wide range of higher order statistics. This section will first explain our overall methodology and then introduce some useful approximations as preliminaries.

#### A. Overall methodology

Fig. 2 illustrates our overall approach to address the connections between spatial and temporal domains. Our analysis implies two underlying assumptions. First, the queueing processes of users are assumed to be independent and homogeneous. This assumption is reasonable because in the macro time-scale, users are assumed to have independent mobility traces; while in the micro time-scale, users are allowed to hop randomly between independent bands. The composite effects of random mobility and band selection renders the queueing process of a user to be independent of others in the long term. In this case, we can consider a typical user with a typical queueing process, at a typical location and associated with a typical BS. A typical user can be understood as an arbitrary user or a randomly selected user. A probability space can also be defined for the typical user for its states of queueing and signal reception, etc. The second assumption is that all BSs constantly transmit with power  $P_n$ . This assumption decouples the interference statistics with user behavior and represents the worst-case interfering scenario. It is reasonable because the combined load of primary and secondary traffic from multiple users is likely to keep BSs in constant transmission.

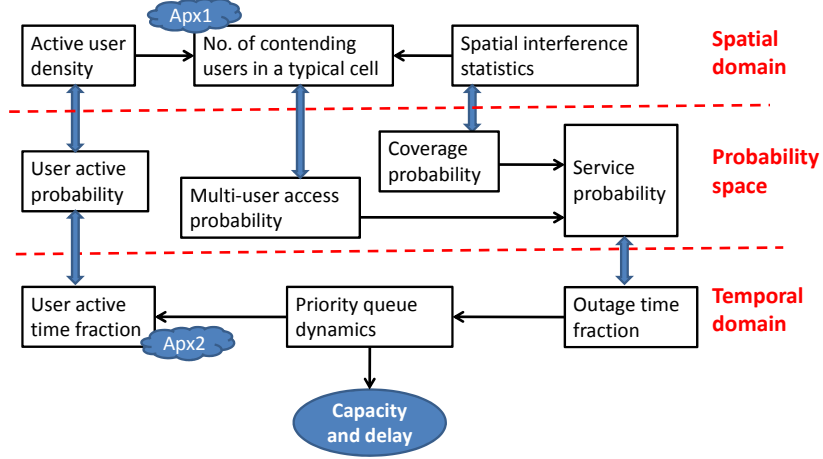


Fig. 2. Connections among parameters in the spatial and temporal domains.

According to the ergodic theory, when the queueing process of the typical user has a statistical equilibrium, the queueing process is ergodic [54] and hence the time average of a queueing parameter is identical to the average over the probability space. This allows us to map time-domain parameters to the probability space. Moreover, according to the theory of Palm probability in stochastic geometry, the spatial average of a large scale network is identical to the probabilistic average over the typical user/BS [49]. This allows us to map spatial-domain parameters to the probability space. Based on these mappings, we are able to deduce a chain of relations in Fig. 2 as follows.

Let us consider the *outage time fraction* in a typical queue, which is the average fraction of time that secondary services is not available. The *outage time fraction* affects the queueing dynamics and hence the *user active time fraction*, which is the average fraction of time that there is secondary traffic buffered in the queue. The *user active time fraction* is identical to the *active probability* of a typical user, which affects the *active user density* in the spatial domain. *Active user density* and *spatial interference statistics* both affect the distribution of the *number of contending users in a typical cell*, which determines the *multi-user access probability*. *Spatial interference statistics* also affects the *coverage probability* of a typical user. Moreover, as shown in (1), the *access probability* and *coverage probability* affects the *service probability*, which ultimately determines the *outage time fraction*. In other words, we have

$$\varepsilon = 1 - \rho_o \quad (9)$$

where  $\varepsilon$  is the service probability,  $\rho_o$  is the outage time fraction, and  $\rho_o$  can be expressed as a function of  $\varepsilon$ . The above chain of relations allows us to establish an equilibrium equation that connects first-order statistics of multiple parameters in the spatial and temporal domains. To establish the equation in an analytical form, two approximations are further introduced.

### B. Approximation to the number of in-coverage users in a typical cell

The probability density function (PDF) of the size of a typical Poisson Voronoi cell is analytically intractable but can be approximated using the Monte Carlo method. Let  $\lambda$  be the density of the underlying Poisson process and  $V$  denote the random size of a typical Voronoi cell normalized by  $1/\lambda$ . The PDF of  $V$  is given by [51]

$$f_V(x) = \frac{3.5^{3.5}}{\Gamma(3.5)} x^{2.5} e^{-3.5x} \quad (10)$$

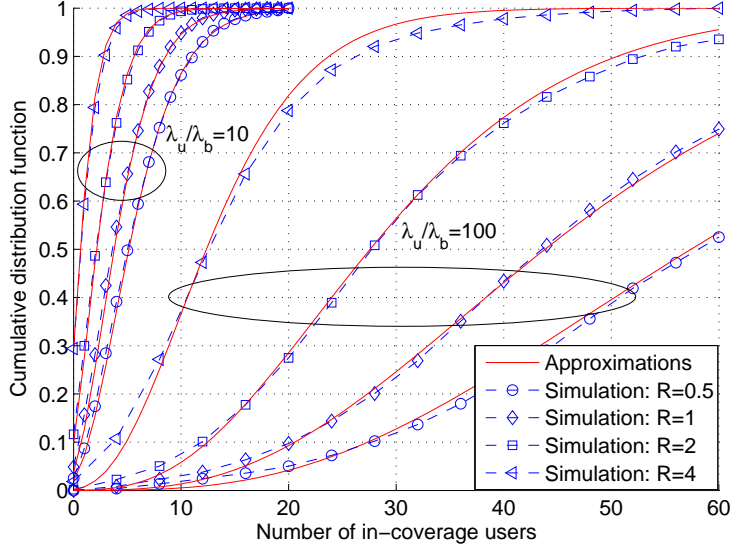


Fig. 3. Approximation on the probability density function of in-coverage users in a typical cell ( $\lambda_b = 10^{-6}$ ).

where  $\Gamma(\cdot)$  is the gamma function. Moreover, consider an arbitrary user and the random size  $U$  of the Voronoi cell to which the user belongs to. The PDF of  $U$  normalized by  $1/\lambda$  is given by [52]

$$f_U(x) = \frac{3.5^{4.5}}{\Gamma(4.5)} x^{3.5} e^{-3.5x}. \quad (11)$$

The difference between  $f_V(x)$  and  $f_U(x)$  comes from the fact that a user has a higher chance to be covered by larger Voronoi cells.

Let us consider a single band of the network with BS density  $\lambda_b$  and user density  $\lambda_u$ . Denoting  $K_1$  as the random number of users in a non-empty Voronoi cell, the probability mass function (PMF) of  $K_1$  is given by

$$f_{K_1}(k) = \int_0^\infty \frac{(\frac{\lambda_u}{\lambda_b}x)^k}{k!} e^{-\frac{\lambda_u}{\lambda_b}x} f_U(x) dx. \quad (12)$$

Let  $K$  be the random number of ‘in-coverage’ users in a Voronoi cell. The distribution of  $K$  is related to the size and shape of the cell and it is difficult to obtain its exact PMF. Keeping the basic form of (12), we propose an approximation to the PMF of  $K$  given by

$$\begin{aligned} f_K(k) &\approx \int_0^\infty \frac{(p\Lambda \frac{\lambda_u}{\lambda_b}x)^k}{k!} e^{-p\Lambda \frac{\lambda_u}{\lambda_b}x} f_U(x) dx \\ &= \frac{3.5^{4.5}\Gamma(4.5+k)}{\Gamma(4.5)k!} \frac{(\Lambda\lambda_u p/\lambda_b)^k}{(3.5 + \Lambda\lambda_u p/\lambda_b)^{4.5+k}}. \end{aligned} \quad (13)$$

where the parameters  $p$  and  $\Lambda$  are introduced to capture the effect of colored thinning on the original user point process. Here,  $p$  is the probability that an arbitrary user falls within coverage (with target rate  $R$ ) and can be calculated by (5). The coefficient  $\Lambda$  is an artificial constant to capture the effect of colored thinning. The value of  $\Lambda$  is obtained by searching for the best fit of (13) to the empirical PMF obtained via Monte Carlo simulations. Through extensive simulations we find that given  $\Lambda = 2/3$ , the approximation in (13) is valid for a wide range of practical values for  $\lambda_u$  and  $\lambda_b$ . Fig. 3 illustrates the accuracy of this approximation.



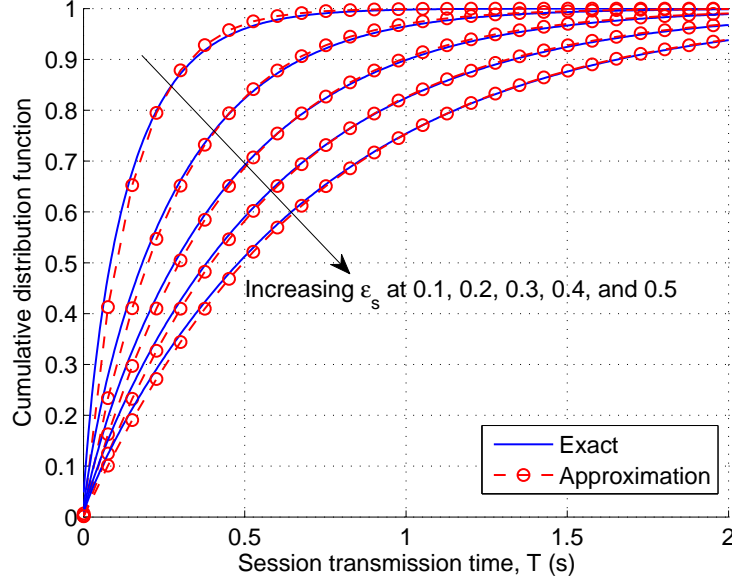


Fig. 4. Exponential approximation for the CDF of session transmission time  $T$  ( $\bar{\alpha}_o = 0.1$ ,  $\bar{\alpha}_s = 1$ ,  $\varepsilon_o = 0.3$ ).

### C. Approximation to user active time fraction

A user is active when there are sessions buffered or being transmitted in the queue. We are interested in the probability  $p_{active}$  that a typical user stays active. This probability also represents the fraction of time for a user to be active. Let  $T$  be the random transmission time of a session. The mean value of  $T$  is given by [54]

$$\bar{T} = \frac{\bar{\beta}_s}{1 - \rho_o}. \quad (14)$$

The exact PDF of  $T$  is not exponential, but for the purpose of calculating the user active probability, we assume that  $T$  follows an exponential distribution with mean  $\bar{T}$ . The accuracy of this approximation is illustrated in Fig. 4, where we assume exponentially distributed  $\beta_o$  and  $\beta_s$ , set  $\bar{\alpha}_o = 0.1$ ,  $\bar{\alpha}_s = 1$ ,  $\varepsilon_o = 0.3$ , and let  $\varepsilon_s$  varies from 0.1 to 0.5. The exact PDF of  $T$  is obtained from its Laplace transform according to (23). We find that the exponential approximation is valid under the condition that the arrival rate of outage is greater than the arrival rate of secondary traffic session. This condition is realistic because our system model considers delay at the session level, which has a larger time scale than outages caused by packet-level primary traffic.

Now let us consider a discrete-value stochastic process representing the number of sessions staying in the queue. Based on the above mentioned exponential approximation, it is easy to see that this process is a classic birth-death process [54] characterized by an uniform birth rate  $1/\bar{\alpha}_s$  and death rate  $1/\bar{T}$ . Let  $\phi_k$  ( $k = 0, 1, 2, 3, \dots$ ) denote the steady state probability that there are  $k$  sessions in the queue. The equilibrium condition of the birth-death process gives  $\phi_k = (\bar{T}/\bar{\alpha}_s)^k \phi_0$ . By further considering the constraint of total probability  $\sum_{k=0}^{\infty} \phi_k = 1$ , we have  $\phi_0 = 1 - \bar{T}/\bar{\alpha}_s$ . It follows that

$$p_{active} = 1 - \phi_0 = \bar{T}/\bar{\alpha}_s = \frac{\bar{\beta}_s}{(1 - \rho_o)\bar{\alpha}_s} = \frac{\rho_s}{1 - \rho_o} = \frac{\rho_s}{\varepsilon}. \quad (15)$$

#### IV. CAPACITY-DELAY TRADEOFF ANALYSIS

##### A. Useful Lemmas

*Lemma 1:* Let  $\varepsilon_n$  denote the probability that a user can be served with a target rate  $R$  by the nearest BS operating in the  $n$ th band. We have

$$\varepsilon_n = \frac{\Omega_n}{\Lambda \lambda_n} \left[ 1 - \left( 1 + \frac{\Lambda p_{n,v} \lambda_n}{3.5} \right)^{-3.5} \right] \quad (16)$$

where  $p_{n,v}$  is given by (5),  $\Lambda = 2/3$ , and

$$\lambda_n = \frac{\lambda_u}{\lambda_{b,n}} \cdot \frac{\rho_s}{\varepsilon} \cdot \frac{\Omega_n p_{n,v}}{\sum_{n=1}^N \Omega_n p_{n,v}}. \quad (17)$$

*Proof:* See Appendix I.  $\square$

*Lemma 2:* When the cellular system independently operates a total number of  $N$  bands, the probability that a user can be served by at least one band with a targeted rate  $R$  is

$$\varepsilon = 1 - \prod_{n=1}^N (1 - \varepsilon_n). \quad (18)$$

*Proof:* It is straightforward to see that  $(1 - \varepsilon)$  equals the joint probability that all bands fail to provide service to a user with the target rate  $R$ .  $\square$

According to Lemma 1,  $\varepsilon_n$  is itself a function of  $\varepsilon$ . Therefore Lemma 2 gives a non-linear equation of  $\varepsilon$ , based on which the value of  $\varepsilon$  can be calculated by solving the non-linear equation via numerical methods. In the special case that all bands have the same characteristics in terms of bandwidth, transmit power, and availability, (18) can be simplified to

$$\varepsilon_N = 1 - (1 - \varepsilon_n)^N. \quad (19)$$

In case when  $N = 1$ , (18) can be solved to give  $\varepsilon$  as an explicit function related to capacity  $C$  and target rate  $R$  as follows

$$\varepsilon_1 = \frac{p_{n,v}}{3.5} \frac{\Lambda \lambda_u}{\lambda_{b,n}} \frac{C}{R} \left[ 1 - \left( 1 - \frac{\Lambda \lambda_u C}{\Omega_n \lambda_{b,n} R} \right)^{-2/7} \right]^{-1}. \quad (20)$$

##### B. General results for capacity-delay tradeoff

Once the value of  $\varepsilon$  is obtained, we can evaluate the mean delay and delay distribution of a session. Established results for two-class M/G/1 priority queues with preemptive-resume policy [54] can be directly applied to give the following two propositions.

*Proposition 1:* The mean delay of a session is given by

$$\bar{D} = \frac{1}{2\varepsilon(\varepsilon - \frac{C}{R})} \left( \frac{\hat{\beta}_s}{\bar{\alpha}_s} + \frac{\hat{\beta}_o}{\bar{\alpha}_o} \right) + \frac{\bar{L}}{R\varepsilon} \quad (21)$$

where  $\hat{\beta}_s$  and  $\hat{\beta}_o$  are the second-order moments of random variables  $\beta_s$  and  $\beta_o$ , respectively.

The delay of a session is the total time the session spends in the queue and consists of two parts. The first part is waiting time  $W$ , which is the duration from the moment of arrival to the moment when the transmission starts. The second part is transmission time  $T$ , which is the duration from the moment when transmission starts to the moment when the transmission ends. It follows that  $D = W + T$ , where  $W$  and  $T$  are independent RVs [54]. The PDF of  $D$  cannot be obtained directly. However, the Laplace transforms of the PDFs of  $W$  and  $T$  can be evaluated. Let  $\mathfrak{L}_X(\cdot)$  denote the Laplace transform to the PDF of random variable  $X$ , we have the following proposition.

*Proposition 2:* The Laplace transform of the random delay  $D$  of a typical session is given by

$$\mathfrak{L}_D(s) = \mathfrak{L}_T(s)\mathfrak{L}_W(s). \quad (22)$$

Here,  $\mathfrak{L}_T(s)$  is given by

$$\mathfrak{L}_T(s) = \mathfrak{L}_{\beta_s} [K(s)] \quad (23)$$

where  $\mathfrak{L}_{\beta_s}(\cdot)$  is the Laplace transform of  $\beta_s$  and

$$K(s) = s + \frac{1 - G(s)}{\bar{\alpha}_o}. \quad (24)$$

Here,  $G(s)$  is the solution with the smallest absolute value that satisfies the following equation

$$x - \mathfrak{L}_{\beta_o} \left( s + \frac{1 - x}{\bar{\alpha}_o} \right) = 0 \quad (25)$$

where  $\mathfrak{L}_{\beta_o}(\cdot)$  is the Laplace transform of  $\beta_o$ . The second term  $\mathfrak{L}_W(s)$  in (22) is given by

$$\mathfrak{L}_W(s) = (1 - \rho_o - \rho_s)\bar{\alpha}_s \frac{K(s)}{\mathfrak{L}_{\beta_s} [K(s)] + \bar{\alpha}_s s - 1}. \quad (26)$$

### C. Capacity-delay tradeoff in special cases

1) *Exponential distribution:* Propositions 1 and 2 are applicable when both the file size  $L$  and outage duration  $\beta_o$  follow general distributions. In the special case where both  $L$  and  $\beta_o$  follow exponential distributions, we have  $\hat{\beta}_s = 2(\bar{\beta}_s)^2$  and  $\hat{\beta}_o = 2(\bar{\beta}_o)^2$ . The mean delay becomes

$$\bar{D} = \frac{1}{\varepsilon(\varepsilon - \frac{C}{R})} \left( \frac{C\bar{L}}{R^2} + (1 - \varepsilon)^2 \bar{\alpha}_o \right) + \frac{\bar{L}}{R\varepsilon}. \quad (27)$$

Moreover, given an exponential random variable  $X \sim \exp(\bar{X})$ , its Laplace transform can be evaluated as

$$\mathfrak{L}_{exp}(s) = \frac{1}{1 + s\bar{X}}. \quad (28)$$

Based on (28), closed-form Laplace transforms of  $\beta_s = L/R$  and  $\beta_o$  can be obtained in (23) and (25). It follows that Eqn. (25) can be solved explicitly to give

$$G(s) = \frac{(1 + \varepsilon_o + s\bar{\beta}_o) - \sqrt{(1 + \varepsilon_o + s\bar{\beta}_o)^2 - 4\varepsilon_o}}{2\varepsilon_o}. \quad (29)$$

2) *Gamma distribution:* A more general distribution we can consider for  $L$  and  $\beta_o$  is Gamma distribution, which provides more flexibility to model a variety of practical scenarios. The PDF of Gamma distribution is given by

$$\Gamma(k, \theta) = \frac{1}{\theta^k} \frac{1}{\Gamma(k)} t^{k-1} e^{-\frac{t}{\theta}} \quad (30)$$

where  $k$  and  $\theta$  are the shape and scale parameters, respectively. The first and second moments of the Gamma distribution are  $k\theta$  and  $k(k+1)\theta^2$ , respectively. Let  $L \sim \Gamma(k_L, \bar{L}/k_L)$  and  $\beta_o \sim \Gamma(k_{\beta_o}, \bar{\beta}_o/k_{\beta_o})$ . Here we introduce two new parameters  $k_L$  and  $k_{\beta_o}$  to characterize the shape of distributions of  $L$  and  $\beta_o$ , respectively. It follows that  $\beta_s = L/R \sim \Gamma(k_L, \bar{L}/(k_L R))$ , and the mean delay in (21) becomes

$$\bar{D} = \frac{1}{2\varepsilon(\varepsilon - \frac{C}{R})} \left( \frac{C\bar{L}}{R^2} \frac{k_L + 1}{k_L} + (1 - \varepsilon)^2 \bar{\alpha}_o \frac{k_{\beta_o} + 1}{k_{\beta_o}} \right) + \frac{\bar{L}}{R\varepsilon}. \quad (31)$$

It is easy to see that when  $k_L = 1$  and  $k_{\beta_o} = 1$ , the Gamma distribution is reduced to exponential distribution and (31) is reduced to (27).

To evaluate the delay distribution, we have the Laplace transform of  $G \sim \Gamma(k, \theta)$  given by

$$\mathfrak{L}_{\text{gamma}}(s) = (1 + \theta s)^{-k}. \quad (32)$$

Based on (32), closed-form Laplace transforms of  $\beta_s = L/R$  and  $\beta_o$  can be obtained according to (23) and (25). It follows that when  $k$  is an integer or a rational fraction, Eqn. (25) yields a polynomial form. Therefore the function  $G(s)$  in (25) can be easily solved using existing root-finding algorithms for polynomials.

## V. CAPACITY LIMIT AND SCALING

This section studies the fundamental capacity limit at the interference limited regime and investigates how the capacity limit scales with bandwidth and user-BS density ratio. The capacity limit is defined as the maximum capacity that permits a stable queue at a typical user. It is also the capacity that gives infinite mean delay. Interference-limited regime means that power  $P_n$  is sufficiently large to justify the closed-form SINR CCDF in (4). For simplicity, we assume that the  $N$  bands have homogeneous characteristics in terms of bandwidth and BS density. Two different cases are considered. The first case assumes a fixed bandwidth of each band, which means the system bandwidth scales linearly with  $N$ . This case is useful when we want to investigate the impact of spectrum aggregation on the system capacity. The second case assumes a fixed system bandwidth, which means the bandwidth per band is inversely proportional to  $N$ . This case is relevant when we are interested in the impacts of spectrum sharing and channelization on the system capacity. Throughout this section, we use the capital letter ‘N’ as the footnote of parameters to emphasize that we consider homogeneous bands. For example,  $W_n$ ,  $\varepsilon_n$  and  $\lambda_n$  are replaced by  $W_N$ ,  $\varepsilon_N$  and  $\lambda_N$ , respectively.

### A. Fixed bandwidth per band

*Proposition 3:* In the case of fixed bandwidth per band, the capacity limit  $C_I^{\text{lim}}$  is a function of  $R$ ,  $\lambda_u$ ,  $\lambda_b$ , and  $N$  given by

$$C_I^{\text{lim}} = R [1 - (1 - \varepsilon_N)^N] \quad (33)$$

where

$$\varepsilon_N = \frac{\Omega_N}{\Lambda \lambda_N} \left[ 1 - \left( 1 + \frac{\Lambda \lambda_N p_N^I}{3.5} \right)^{-3.5} \right]. \quad (34)$$

Here,  $p_N^I$  is given by

$$p_N^I = \left( 1 + \sqrt{2^{R/W_N} - 1} \arctan \sqrt{2^{R/W_N} - 1} \right)^{-1} \quad (35)$$

and  $\lambda_N = \lambda_u / (\lambda_{b,n} N)$ .

*Proof:* A stable queue requires  $1 - \rho_o - \rho_s > 0$ , which gives  $\varepsilon > \rho_s = C/R$ . The capacity limit is achieved when the equality holds, i.e.,  $\varepsilon = C/R$  or  $\rho_d/(1 - \rho_o) = 1$ . Substituting this equation into Lemma 1 yields  $\varepsilon_N$  in (34).  $\square$

We note that by considering the limiting condition,  $\varepsilon_N$  can be expressed as an explicit function of other parameters (as opposed to numerically solving a non-linear equation in Lemma 2). This allows us to express the capacity limit as a closed-form function of  $R$ ,  $N$ ,  $\lambda_u$ , and  $\lambda_b$ , as shown in (33). In the case of fixed bandwidth per band, we are interested in the following optimization problem: given  $N$  and the network environment  $\lambda_u$  and  $\lambda_b$ , how can we choose a proper target rate  $R$  to maximize the capacity limit? This optimization problem can be formally stated as  $C_I^{\text{max}} = \max_R(C_I^{\text{lim}})$ . To better understand the nature of this optimization problem, representative numerical examples are presented in Fig. 5 to show  $C_I^{\text{lim}}$  as a function of  $R$ . We see that there is a unique maximum value of  $C_I^{\text{lim}}$ , which is achieved when the first-order derivative  $dC_I^{\text{lim}}/dR$  equals zero. According to Proposition 3, the derivative function  $dC_I^{\text{lim}}/dR$  can be obtained in closed-form to give the following corollary.

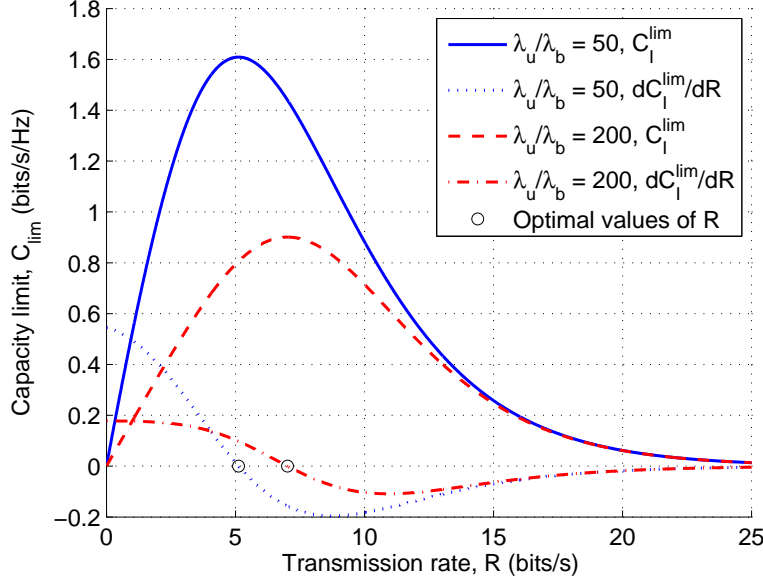


Fig. 5. Capacity limit  $C_I^{\text{lim}}$  and its first-order derivative as a function of  $R$  (fixed bandwidth per band,  $N=5$ ).

*Corollary 1:* The optimum value  $R$  for the optimization problem  $C_I^{\text{max}} = \max_R(C_I^{\text{lim}})$  is given by the root of the following non-linear equation:

$$\frac{dC_I^{\text{lim}}}{dR} = Rf'_o(R) + f_0(R) = 0 \quad (36)$$

where

$$f_0(R) = 1 - (1 - \varepsilon_N)^N \quad (37)$$

$$f'_o(R) = f_1(R) \cdot f_2(R) \cdot f_3(R) \cdot f_4(R) \quad (38)$$

$$f_1(R) = N(1 - \varepsilon_N)^{N-1} \quad (39)$$

$$f_2(R) = \left(1 + \frac{\lambda_N p_N^I}{3.5}\right)^{-4.5} \quad (40)$$

$$f_3(R) = -\frac{\arctan(\chi_N) + (1 + \chi_N^2)^{-1}}{(1 + \chi_N \arctan(\chi_N))^2} \quad (41)$$

$$f_4(R) = \frac{\ln 2}{2} 2^R (2^R - 1)^{-1/2} \quad (42)$$

$$\chi_N = \sqrt{2^R - 1}. \quad (43)$$

In the above equations,  $p_N^I$  is defined in (35) and  $\varepsilon_N$  is defined in (34).

Based on the above corollary, the first-order derivative function  $dC_I^{\text{lim}}/dR$  is calculated and shown in Fig. 5. The root is obtained by solving the non-linear equation and shown to be accurate for achieving the maximum value of  $C_I^{\text{lim}}$ .

### B. Fixed system bandwidth

In this case, the total system bandwidth is normalized to 1 and the bandwidth of each band becomes  $1/N$ . Define the capacity limit  $C_{II}^{\text{lim}}$  as the maximum achievable capacity for a stable queue given  $R$ ,

$N$ ,  $\lambda_u$ , and  $\lambda_b$ . Further define the maximum capacity as  $C_{II}^{\max} = \max_R(C_{II}^{\lim})$ . We have the following two propositions.

*Proposition 4:* The capacity limit  $C_{II}^{\lim}$  can be calculated according to Proposition 3 by replacing  $p_N^I$  with  $p_N^{II}$ , where

$$p_N^{II} = \left(1 + \sqrt{2^{RN/W_N} - 1} \arctan \sqrt{2^{RN/W_N} - 1}\right)^{-1}. \quad (44)$$

*Proof:* The proof is straightforward by following the proof of Proposition 3 and setting the channel bandwidth to  $1/N$ .  $\square$

*Proposition 5:* The maximum capacity is given by

$$C_{II}^{\max} = C_I^{\max}/N \quad (45)$$

where  $C_I^{\max}$  can be calculated from Corollary 1.

*Proof:* According to Propositions 3 and 4, we can write  $C_{II}^{\lim}(R) = C_I^{\lim}(RN)/N$ . Further considering the fact that adding a scaling on  $R$  will not change the maximum value of  $C_I^{\lim}$ , i.e.,  $\max_R C_I^{\lim}(R) = \max_R C_I^{\lim}(RN) = C_I^{\max}$ , Proposition 5 can be proved.  $\square$

## VI. NUMERICAL RESULTS AND DISCUSSIONS

This section presents numerical results and discusses their implications. First, we aim to understand the impacts of various parameters on the capacity-delay tradeoff (Figs. 6 to 10). Second, we want to investigate how the fundamental capacity limit scales with the number of bands  $N$  and user-BS density ratio (Figs. 11 to 13). For illustration purpose, we consider an interference-limited system and homogeneous bands with  $W_N = 1$  and  $\Omega_N = 1$ .

### A. Capacity-delay tradeoff

Due to page limits, we restrict our discussions to the mean delay and the case of fixed bandwidth per band. Except when otherwise mentioned, the default parameter values are set to be  $N = 5$ ,  $\lambda_u/\lambda_b = 50$ ,  $\bar{L} = 10$ , and  $\bar{\alpha}_o = 10$ . Moreover, the distributions of  $L$  and  $\alpha_o$  are treated as exponential. Therefore, our subsequent discussions are primarily based on Eqn. (27).

Fig. 6 shows the mean delay  $\bar{D}$  as a function of  $R$  with varying  $N$  while the capacity is fixed to  $C = 1$  bits/s. U-shape curves are observed, indicating that given other parameters, there is an optimal value for  $R$  to minimize the mean delay. Because we are interested in the fundamental capacity-delay tradeoff, it is desirable to consider the minimized delay over feasible values of  $R$ . Define  $\bar{D}_{\min} = \min_R(\bar{D})$ , we will subsequently evaluate  $\bar{D}_{\min}$  as a function of  $C$ . The value of  $\bar{D}_{\min}$  is obtained by performing a numerical optimization over  $R$ .

Fig. 7 shows the impact of  $\lambda_u/\lambda_b$  on the capacity-delay tradeoff curve. Two interesting phenomena are observed. First, when the user-BS density ratio is relatively high ( $100 \leq \lambda_u/\lambda_b \leq 1000$ ), the capacity per user (at a fixed delay) appears to scale linearly with  $\lambda_b/\lambda_u$ . We called this “infrastructure-limited” regime, in which the investment in BS infrastructure yields linear returns on the capacity. In contrast, when the user-BS density is relatively low ( $10 \leq \lambda_u/\lambda_b \leq 100$ ), investment in BS infrastructure only yields sub-linear returns. Second, in the low delay regime, there is minimum delay even when  $C$  approaches zero. Such a minimum delay is caused by coverage outage and primary traffic interruption, which caps the secondary service probability.

Fig. 8 shows the impact of the number of channels  $N$  on the capacity-delay tradeoff curve. The capacity limits with respect to different values of  $N$  are also shown. The delays are shown to rise quickly when  $C$  approaches the capacity limits. It is observed that in the medium to high delay regime, capacity at a fixed delay scales linearly with  $N$ . In the low delay regime, increasing  $N$  contributes slightly to reducing the minimum delay. Fig. 8 indicates that spectrum aggregation is effective for both capacity enhancement and delay reduction.

Fig. 9 shows the impact of average file size  $\bar{L}$  on the capacity-delay tradeoff curve. The capacity limit is also shown, which is unrelated to the value of  $\bar{L}$ . In the low to medium capacity regime,  $\bar{L}$  is shown to have a significant effect on the delay. A smaller value of  $\bar{L}$  leads to a smaller delay because the file transmission has a lower probability of being interrupted by an outage. In the high delay regime, the impact of  $\bar{L}$  diminishes as all delay curves eventually converge to the capacity limit. Fig. 9 suggests that file/session size management is an important factor to consider if a system is designed for low delay performance.

Fig. 10 shows the impact of mean outage arrival interval  $\bar{\alpha}_o$  on the capacity-delay tradeoff curve. The capacity limit, which is independent from the values of  $\bar{\alpha}_o$ , is also shown. In the low delay regime, the curves converge to a minimum delay. In the high delay regime, we can predict that the curves also slowly converge to the capacity limit. However, significant differences are observed in the low to medium delay regimes. A smaller value of  $\bar{\alpha}_o$  leads to smaller delays. This is because an interrupted session is less likely to be prolonged for a long period. Fig. 10 implies that introducing extra dynamics into the system (such as dynamic scheduling) can potentially help to reduce the delay.

### B. Capacity limit and scaling

This subsection investigates how the capacity limit scales with  $N$  and user-BS density ratio. Consider the case of fixed bandwidth per band, Fig. 11 applies Corollary 1 to show the maximum capacity  $C_I^{\max}$  as a function of  $N$  with varying  $\lambda_u/\lambda_b$ . We see that the capacity increases monotonically with increasing  $N$ , indicating the benefits of spectrum aggregation. However, increasing  $N$  shows diminishing returns on the capacity gain. This differs from the intuition that system capacity scales linearly with bandwidth (i.e., the number of bands). It is also interesting to observe that the curves with different values of  $\lambda_u/\lambda_b$  converge to the same value when  $N$  tends large. This is because we assume that a user is allowed to access only one band. When  $N$  is large, the capacity is limited by the bandwidth per band rather than the number of bands. Fig. 11 suggests that to achieve the full potential of spectrum aggregation, it is important to allow users to access multiple bands simultaneously.

Considering the case of fixed system bandwidth, Fig. 12 applies Proposition 5 to show the maximum capacity  $C_{II}^{\max}$  as a function of  $N$  with varying  $\lambda_u/\lambda_b$ . It is shown that with increasing  $N$ , the capacity increases initially but eventually declines. For each value of  $\lambda_u/\lambda_b$ , there exists an optimal value of  $N$  to maximize the capacity. Fig. 12 reveals a design tradeoff between maximizing single channel capacity and maximizing multi-user access probability. When  $N$  increases, the single channel capacity decreases due to reduced bandwidth, while the access probability increases due to increased number of bands. Fig. 12 implies that proper channelization of the available spectrum resource is important, particularly when  $\lambda_u/\lambda_b$  is small.

By performing a numerical search for the optimal value of  $N$  based on results in Fig. 12, Fig. 13 shows the corresponding maximum values of  $C_{II}^{\max}$  as a function of  $\lambda_u/\lambda_b$ . We find that there exists a convenient approximation given by

$$C_{II}^{*\max} \approx 0.6359 - 0.052 \log_2(\lambda_u/\lambda_b). \quad (46)$$

The actual values obtained from numerical calculation and the approximated values obtained from (46) are compared in Fig. 13. It is shown that our approximation is reasonably accurate for  $2 < \lambda_u/\lambda_b < 500$ . Fig. 13 shows that per user throughput is upper bounded by a constant and reduces at a sub-linear rate with increasing  $\lambda_u/\lambda_b$ .

## VII. CONCLUSIONS

An analytical framework has been proposed for the study of the capacity-delay tradeoff in cellular networks with spectrum aggregation. The framework compliments existing ones by focusing on the secondary traffic and offering tractable analytical insights. Analytical results have been derived to characterize the capacity-delay tradeoff and the fundamental capacity limit. Numerical studies have shown that while

spectrum aggregation primarily affects the capacity in the high-delay regime, session size management and dynamic scheduling have bigger impacts on the capacity in the low delay regime. Moreover, when different bands have homogeneous configurations, it has been shown that the per user throughput per Hertz is upper bounded by a constant and reduces at a rate proportional to the logarithm of user-BS density ratio. Our analysis offers useful guidelines for providing novel secondary services over cellular networks to improve the overall capacity utilization.

## APPENDIX I

This appendix gives the proof for Lemma 1. Because we assume that an active user randomly selects a band for access, in an equilibrium state, the density of users in a band is proportional to the area fraction of coverage of this band. The density of active users in the  $n$ th band is then given by

$$\lambda_{u,n} = \lambda_u \cdot p_{active} \cdot \frac{\Omega_n p_{n,v}}{\sum_{n=1}^N \Omega_n p_{n,v}} = \frac{\lambda_u \rho_s}{\varepsilon} \cdot \frac{\Omega_n p_{n,v}}{\sum_{n=1}^N \Omega_n p_{n,v}}. \quad (47)$$

Now consider an active user in band  $n$ , the number of contenting users in the same cell can be evaluated according to (13) with user density  $\lambda_{u,n}$  and BS density  $\lambda_{b,n}$ . When strict fairness is assumed, the access probability of a user is given by

$$\begin{aligned} p_{n,a} &= \sum_{k=0}^{\infty} \frac{1}{k+1} f_K(k) \\ &= \sum_{k=0}^{\infty} \frac{1}{k+1} \int_0^{\infty} \frac{(\lambda_n \Lambda p_{n,v} x)^k}{k!} e^{-\lambda_n \Lambda p_{n,v} x} f_U(x) dx \\ &= \sum_{k=0}^{\infty} \frac{1}{\lambda_n \Lambda p_{n,v} x} \left[ \int_0^{\infty} \frac{(\lambda_n \Lambda p_{n,v} x)^k}{k!} \right] e^{-\lambda_n \Lambda p_{n,v} x} f_U(x) dx \\ &= \sum_{k=0}^{\infty} \frac{1}{\lambda_n \Lambda p_{n,v} x} (1 - e^{-\lambda_n \Lambda p_{n,v} x}) f_U(x) dx \\ &= \frac{3.5^{4.5}}{\Gamma(4.5)} \frac{1}{\lambda_n \Lambda p_{n,v}} \left[ \int_0^{\infty} x^{2.5} e^{-3.5x} dx - \int_0^{\infty} x^{2.5} e^{(-3.5 + \lambda_n \Lambda p_{n,v})x} dx \right] \\ &= \frac{3.5^{4.5}}{\Gamma(4.5)} \frac{1}{\lambda_n \Lambda p_{n,v}} \left[ \frac{\Gamma(3.5)}{3.5^{3.5}} - \frac{\Gamma(3.5)}{(3.5 + \lambda_n \Lambda p_{n,v})^{3.5}} \right] \\ &= \frac{1}{\lambda_n \Lambda p_{n,v}} \left[ 1 - \left( 1 + \frac{\Lambda p_{n,v} \lambda_n}{3.5} \right)^{-3.5} \right]. \end{aligned} \quad (48)$$

Finally, Lemma 1 can be obtained by substituting (48) into (1).

## REFERENCES

- [1] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial modeling of the traffic density in cellular networks," *IEEE Wireless Commun. Mag.*, vol. 21, no. 1, pp. 80-88, Feb. 2014.
- [2] "Cisco visual networking index: Global mobile data traffic forecast update 2015-2020," Cisco White Paper, Feb. 2016.
- [3] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "Green delivery: proactive content caching and push with energy-harvesting-based small cells," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142-149, Apr. 2015.
- [4] X. Wang, M. Chen, T. Taleb, A. Ksentini and V. C. M. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131-139, Feb. 2014.
- [5] N. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Info. Theory*, vol. 58, no. 10, pp. 6524-6540, Oct. 2012.
- [6] Y. Zhang, H. Lu, H. Wang, and X. Hong, "Cognitive cellular content delivery networks: cross-layer design and analysis," *Proc. VTC-Spring'16*, May 2016, Nanjing, China.
- [7] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 122-130, Feb. 2014.



- [8] X. Ge, S. Tu, T. Han, Q. Li, and G. Mao, "Energy efficiency of small cell backhaul networks based on Gauss-Markov mobile models," *IEET Netw.*, vol. 4, no. 2, pp. 158–167, Mar. 2015.
- [9] H.-S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: a comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.
- [10] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.
- [11] W. C. Cheung, T. Q. S. Quek, and M. Kountouris, "Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 561–574, Apr. 2012.
- [12] M. Di Renzo, A. Guidotti, and G. E. Corazza, "Average rate of downlink heterogeneous cellular networks over generalized fading channels: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 3050–3071, Jul. 2013.
- [13] A. Guo and M. Haenggi, "Spatial stochastic models and metrics for the structure of base stations in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5800–5812, Nov. 2013.
- [14] Y. S. Soh, T. Q. S. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 840–850, May 2013.
- [15] X. Lin, J. G. Andrews, and A. Ghosh, "Modeling, analysis and design for carrier aggregation in heterogeneous cellular networks," *IEEE Trans. Commun.*, vol. 61, no. 9, pp. 4002–4015, Sep. 2013.
- [16] A. K. Gupta, H. S. Dhillon, S. Vishwanath, and J. G. Andrews, "Downlink multi-antenna heterogeneous cellular network with load balancing," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4052–4067, Nov. 2014.
- [17] Z. Zeinalpour-Yazdi and S. Jalali, "Outage analysis of uplink two-tier networks," *IEEE Trans. Commun.*, vol. 62, no. 9, pp. 3351–3362, Sep. 2014.
- [18] G. Nigam, P. Minero, and M. Haenggi, "Coordinated multipoint joint transmission in heterogeneous networks," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4134–4146, Nov. 2014.
- [19] X. Zhang and M. Haenggi, "A stochastic geometry analysis of inter-cell interference coordination and intra-cell diversity," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6655–6669, Dec. 2014.
- [20] H. S. Dhillon, Y. Li, P. Nuggehalli, Z. Pi, and J. G. Andrews, "Fundamentals of heterogeneous cellular networks with energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2782–2797, May 2014.
- [21] N. Deng, W. Zhou, and M. Haenggi, "The Ginibre point process as a model for wireless networks with repulsion," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 107–121, Jan. 2015.
- [22] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 4, pp. 301–313, Oct. 1999.
- [23] S. B. Fredj, T. Bonald, A. Proutiere, G. Regnie, and J. W. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level," *SIGCOMM'01*, vol. 31, no. 4, Aug. 2001, pp. 111–122.
- [24] Q. Liu, S. Zhou, and G. Giannakis, "Queueing with adaptive modulation and coding over wireless links: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1142–1153, May 2005.
- [25] L. Le, E. Hossain, and A. Alfa, "Delay statistics and throughput performance for multi-rate wireless networks under multiuser diversity," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3234–3243, Nov. 2006.
- [26] L. Jiao, F. Li, and V. Pla, "Modeling and performance analysis of channel assembling in multichannel cognitive radio networks with spectrum adaptation," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2686–2697, Jul. 2012.
- [27] V. Tumuluru, P. Wang, D. Niyato, and W. Song, "Performance analysis of cognitive radio spectrum access with prioritized traffic," *IEEE Trans. Veh. Technol.*, vol. 61, no. 4, pp. 1895–1906, May 2012.
- [28] J. Wang, A. Huang, W. Wang, and T. Quek, "Admission control in cognitive radio networks with finite queue and user impatience," *IEEE Wireless Commun. Lett.*, vol. 2, no. 2, pp. 175–178, Apr. 2013.
- [29] V. Tumuluru, P. Wang, and D. Niyato, "A novel spectrum scheduling scheme for multichannel cognitive radio network and performance analysis," *IEEE Trans. Veh. Technol.*, vol. 60, no. 4, pp. 1849–1858, May 2011.
- [30] M. Rashid, M. Hossain, E. Hossain, and V. Bhargava, "Opportunistic spectrum scheduling for multiuser cognitive radio: a queueing analysis," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 5259–5269, Oct. 2009.
- [31] N.-S. Vo, T. Duong, H.-J. Zepernick, and M. Fiedler, "A cross-layer optimized scheme and its application in mobile multimedia networks with qos provision," *IEEE Syst. J.*, vol. PP, no. 99, pp. 1–14, 2015.
- [32] S. Lirio Castellanos-Lopez, F. Cruz-Perez, M. Rivero-Angeles, and G. Hernandez-Valdez, "Joint connection level and packet level analysis of cognitive radio networks with VoIP traffic," *IEEE J. Select. Areas Commun.*, vol. 32, no. 3, pp. 601–614, Mar. 2014.
- [33] M. Haenggi, "The local delay in Poisson networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1788–1802, Mar. 2013.
- [34] Z. Gong and M. Haenggi, "The local delay in mobile Poisson networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4766–4777, Sep. 2013.
- [35] M. Neely and E. Modiano, "Capacity and delay tradeoffs for ad-hoc mobile networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 1917–1937, June 2005.
- [36] A. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-Off in wireless networks Part I: The fluid model," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2568–2592, Jun. 2006.
- [37] A. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-delay trade-off in wireless networks part II: Constant-size packets," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5111–5116, Nov. 2006.
- [38] P. Li, C. Zhang, and Y. Fang, "Capacity and delay of hybrid wireless broadband access networks," *IEEE J. Sel. Area, Commun.*, vol. 27, no. 2, pp. 117–125, Feb. 2009.
- [39] X. Ta, G. Mao, and B.D.O. Anderson, "On the giant component of wireless multihop networks in the presence of shadowing," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 5152–5163, Nov. 2009.
- [40] A. J. Fehske and G. P. Fettweis, "Aggregation of variables in load models for cellular data networks," *Proceedings ICC 2012*, Ottawa, Canada, May 2012, pp. 5102–5107.

- [41] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, June 2012.
- [42] A. J. Fehske and G. P. Fettweis, "On flow level modeling of multi-cell wireless networks," *Proc. IEEE 11th Int. Model. Optim. Mobile Ad Hoc Wireless Netw.*, Tsukuba, Japan, May 2013, pp. 572–579.
- [43] I.-H. Hou, V. Borkar, and P. R. Kumar, "A theory of QoS for wireless," *Proc. IEEE INFOCOM*, Rio de Janeiro, Brazil, Apr. 2009, pp. 486–494.
- [44] S. Lashgari and A. S. Avestimehr, "Timely throughput of heterogeneous wireless networks: Fundamental limits and algorithms," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8414–8433, Dec. 2013.
- [45] G. Zhang, T. Q. S. Quek, A. Huang and H. Shan, "Delay and reliability tradeoffs in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1101–1113, Feb. 2016.
- [46] B. Baszczyszyn, M. Jovanovic and M. K. Karray, "Performance laws of large heterogeneous cellular networks," *Proc. 13th Int'l Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, Mumbai, 2015, pp. 597–604.
- [47] H. Ye, C. Liu, X. Hong, and H. Shi, "Uplink capacity-delay trade-off in hybrid cellular D2D networks with user collaboration," *Proc. IEEE WPMC*, Nov. 2016, Shenzhen, China.
- [48] L. Chen, W. Luo, C. Liu, X. Hong, and J. Shi, "Capacity-delay trade-off in collaborative hybrid ad-hoc networks with coverage sensing," *MDPI Sensors*, under review.
- [49] D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*, 2nd Edition, Wiley, 2008.
- [50] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
- [51] J.-S. Ferenc and Z. Neda, "On the size distribution of Poisson Voronoi cells," *Physica A: Statistical Mechanics and its Applications*, vol. 385, no. 2, pp. 518–526, 2007.
- [52] S. M. Yu and S.-L. Kim, "Downlink capacity and base station density in cellular networks," *Proc. WiOpt 2013*, Tsukuba Science City, 2013, pp. 119–124.
- [53] X. Hong, Y. Jie, C. X. Wang, J. Shi and X. Ge, "Energy-spectral efficiency trade-off in virtual MIMO cellular systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2128–2140, Oct. 2013.
- [54] J. W. Cohen, *The Single Server Queue*, North-Holland Publishing Company, 1982.

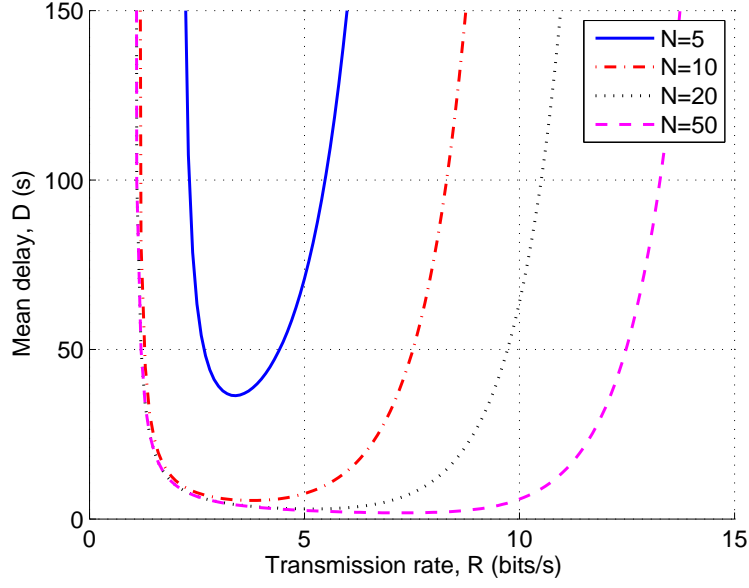


Fig. 6. Mean delay  $\bar{D}$  as a function of  $R$  with varying  $N$  (fixed bandwidth per band,  $C=1$  bit/s/Hz,  $\lambda_u/\lambda_b=50$ ,  $\bar{L}=10$ ,  $\bar{\alpha}_o=10$ ).

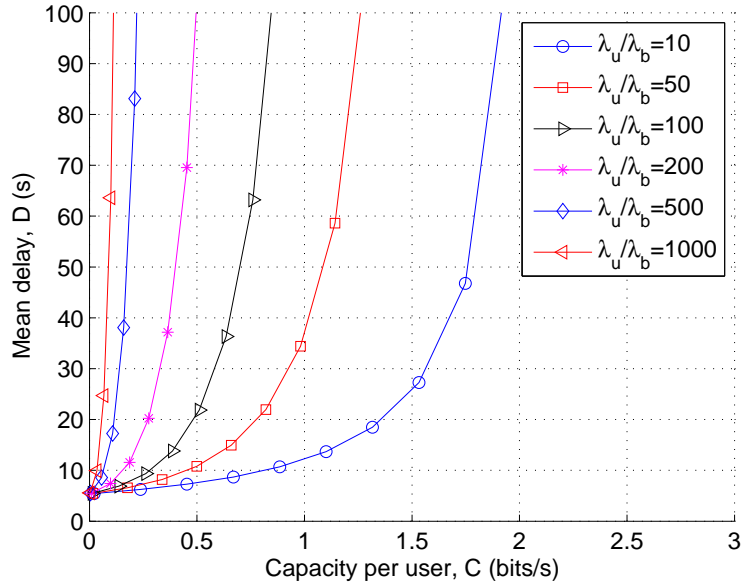


Fig. 7. Mean delay  $\bar{D}$  as a function of per user capacity  $C$  with varying  $\lambda_u/\lambda_b$  (fixed bandwidth per band,  $N=5$ ,  $\bar{L}=10$ ,  $\bar{\alpha}_o=10$ ).

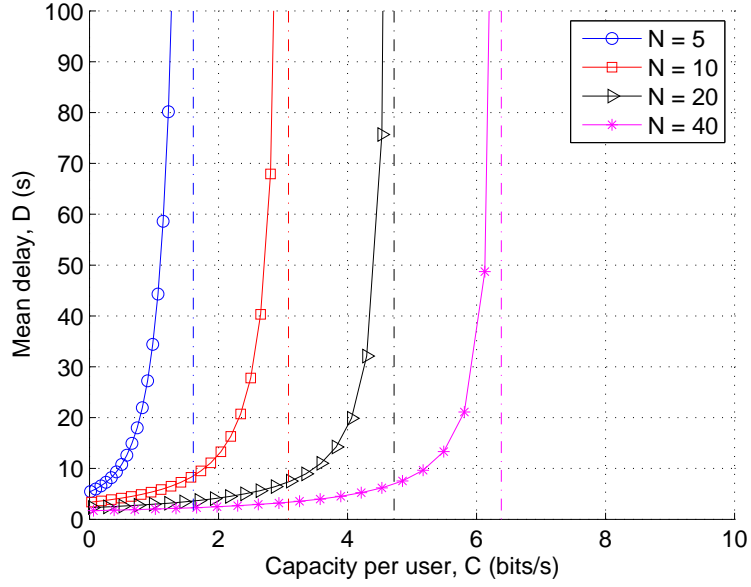


Fig. 8. Mean delay  $\bar{D}$  as a function of per user capacity  $C$  with varying  $N$  (fixed bandwidth per band,  $\lambda_u/\lambda_b=50$ ,  $\bar{L}=10$ ,  $\bar{\alpha}_o=10$ ).

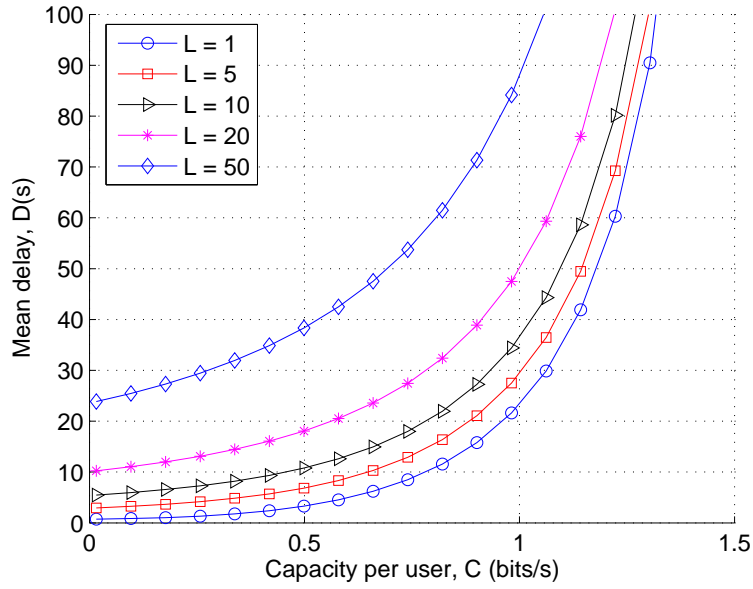


Fig. 9. Mean delay  $\bar{D}$  as a function of per user capacity  $C$  with varying  $\bar{L}$  (fixed bandwidth per band,  $\lambda_u/\lambda_b=50$ ,  $N=5$ ,  $\bar{\alpha}_o=10$ ).

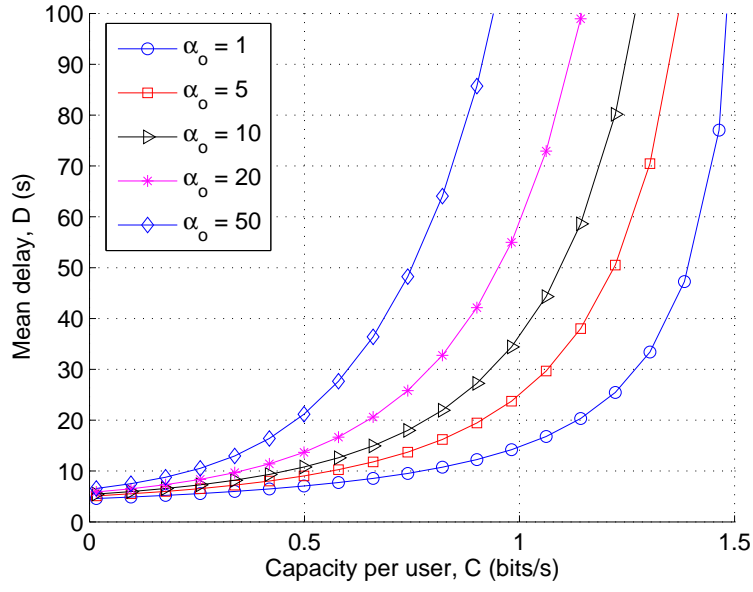


Fig. 10. Mean delay  $\bar{D}$  as a function of per user capacity  $C$  with varying  $\bar{\alpha}_o$  (fixed bandwidth per band,  $\lambda_u/\lambda_b=50$ ,  $N=5$ ,  $\bar{L}=10$ ).

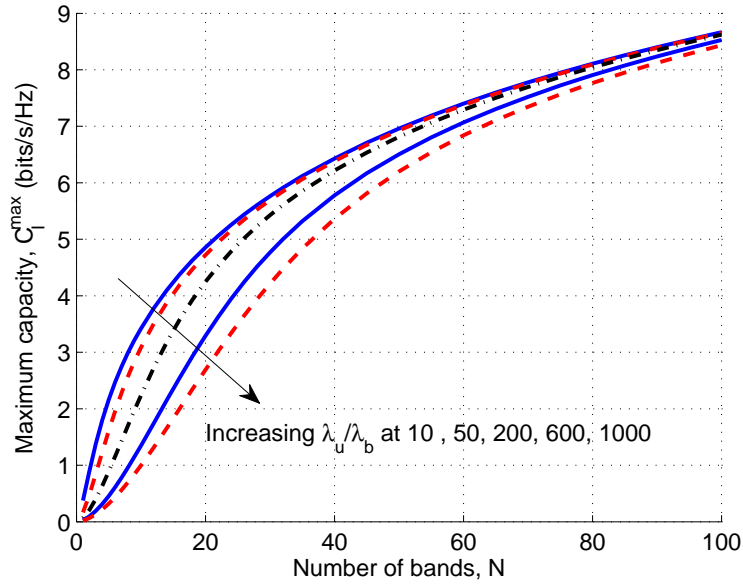


Fig. 11. Maximum capacity  $C_I^{\max}$  as a function of  $N$  with varying  $\lambda_u/\lambda_b$  (fixed bandwidth per band).

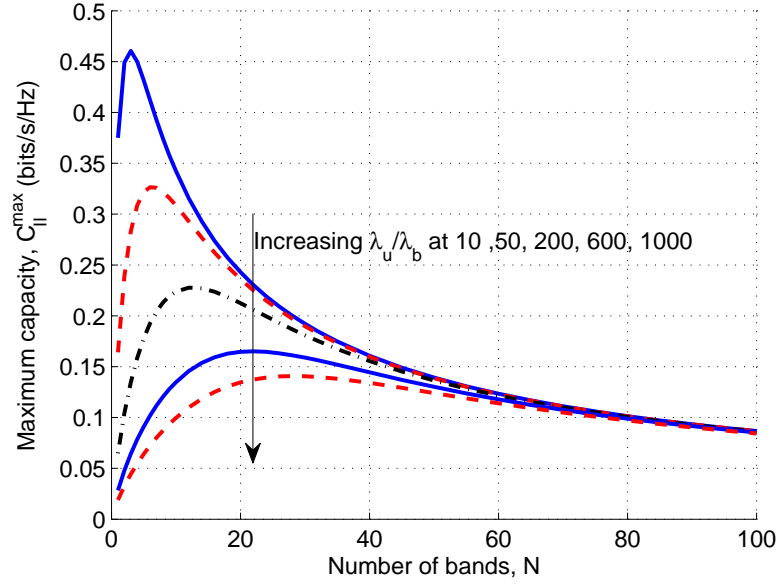


Fig. 12. Maximum capacity  $C_{II}^{\max}$  as a function of  $N$  with varying  $\lambda_u/\lambda_b$  (fixed system bandwidth).

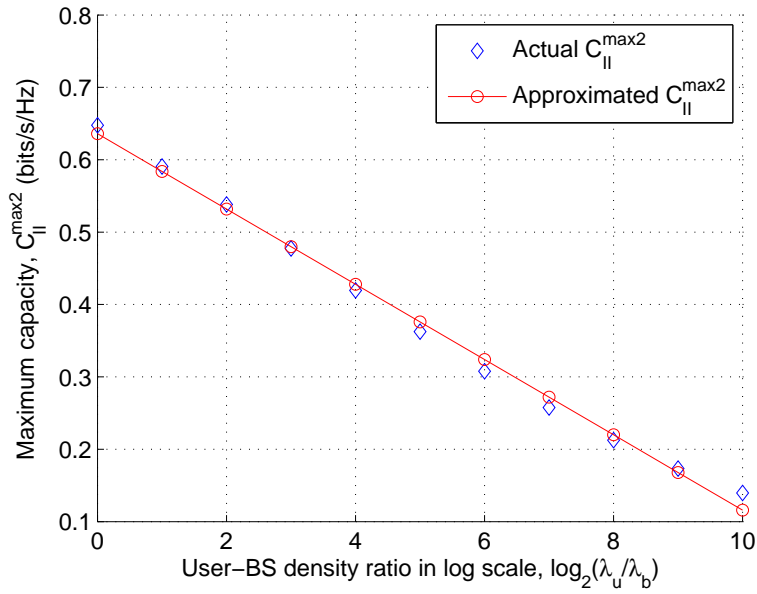


Fig. 13. Approximation of  $C_{II}^{\max}$  as a function of user-BS density ratio (fixed system bandwidth).